Questions for Data Transformation & retrieving in cloud based platforms

Part 1 – Data task

The attached file (file-name) contains a sample of XYZ data related to a CBT operation, representing a beneficiary dataset from Country Office x. The dataset shows beneficiaries enrolled in different activities.

The file contains the following columns:

a. household_identifier: the identifier of the household receiving assistance from XYZ. A household can have x number of beneficiaries associated with it.
b. beneficiary_identifier: the identifier of the beneficiary receiving assistance from XYZ. Individual beneficiaries are associated with households.
c. activity_name: the name of the activity to which a beneficiary is enrolled to receive assistance. Different activities respond to different needs of beneficiary populations.
d. date_of_birth: the date of birth of the beneficiary receiving assistance.
e. gender: gender of the beneficiary

Based on the dataset, please respond to the following:

A Senior Manager working on the design of CBT programmes would like to understand the following:

a. **demographic profile (number of household members, age distribution, gender) of beneficiaries for each activity.**
b. **the extent to which beneficiaries appear in more than activity.**

Using Python or R, please create visualizations that respond to the Senior Manager's questions and comment on the results displayed in your visualizations, explaining to a business audience how to interpret the information and making some observations about what it is showing.

Your submission can be in .pdf or .doc or markdown and must contain 1) the python/R code used, 2) visualizations, and 3) commentary.

- Must have: visual output from python/R; business-friendly interpretation of results.
- Nice to have: correct diagram results; highlighting outliers/missing data values; suggested interpretation of outliers (e.g. dirty data or fraud).
- Evaluation criteria: correctness of the diagram; R/python code quality and clarity; clarity of explanation to a business audience.

Part 2: Written responses

1. **How would you describe to a senior XYZ manager the advantages of investing in data analytics to improve accountability to beneficiaries and donors, and to improve effectiveness (including financial benefits) of CBT programmes?**
    - Score: 20%
    - Time: 25'
    - Must have: minimum 1000 characters (lower would imply poor ideas or poor ability to explain and persuade); must include financial benefits; must include explanation of assurance benefits.
    - Nice to have: creative ideas about the benefits (with financial implications) that data science can provide along the cycle of CBT programmes, from analysis of registered beneficiary data to review of transfers to financial providers and redemptions by beneficiaries.
    - Evaluation: based on the professionalism of the explanation to a senior audience, on the relevance of data science concepts used to make the point, and on the ability to think about business relevant aspects (financial implications, assurance).

2. **The beneficiary registration through several digital platforms and by different actors poses several challenges and may lead duplicate identities being registered to receive assistance. What analytics techniques would you recommend adopting to address beneficiary data deduplication?**
    - Must have: must include at least 2 different approaches to deduplication.
    - Nice to have: identification of different techniques for deduplication, mentioning issues related to data quality and to datasets with non-Latin characters.
    - Evaluation: based on the professionalism and pragmatism of the suggested approaches.

3. **When granted access to beneficiary bank accounts, XYZ has the possibility to analyze the redemption data by beneficiaries (ATM and/or POS). In order to detect possible fraud, what**

types of anomalies would you suggest looking for and through which techniques? And how would you recommend defining an approach which is standard, but at the same time considers the specificity of each country's context?

- Must have: must include at least 3 anomaly types.
- Nice to have: more than 5 anomaly types; proposal of parameter-based approaches to be discussed with business users, in order to find the right thresholds to identify what really constitutes an anomaly in each different context.
- Evaluation: based on the quantity and business-relevance of anomaly types identified for beneficiary redemption transactions, as well as on the quality of the explanations of the suggested data-science techniques.